P1.DPR Over Loop Accelerators

Nuno Paulino
CTM
INESC TEC

INESCTEC
TECNOLOGIA E CIÊNCIA
LABORATÓRIO ASSOCIADO
COORDENADO POR INESCPORTO

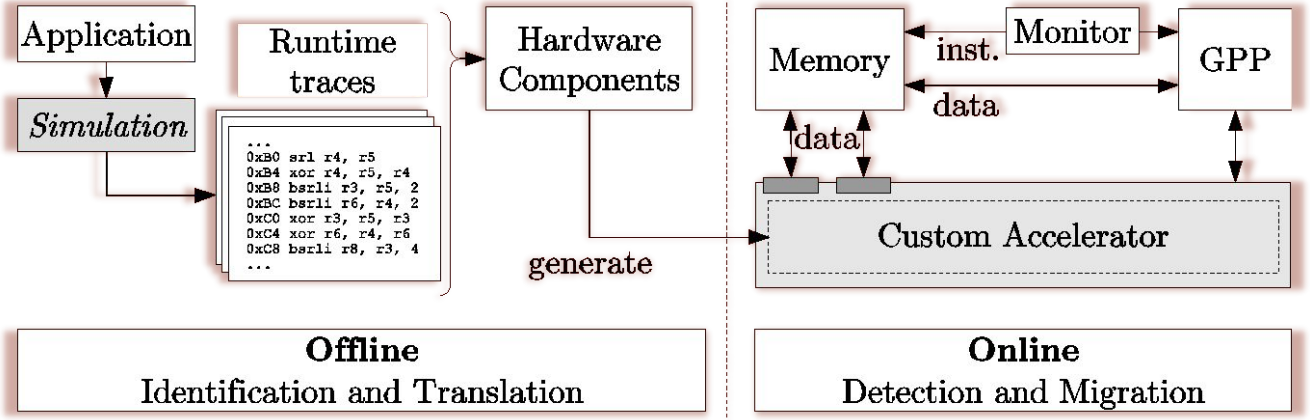
PEPCC

Overview

- Title: *Dynamic Partial Reconfiguration of Customized Single-Row Accelerators, IEEE VLSI, 2019, (https://ieeexplore.ieee.org/document/8502926)*
- Journal: *IEEE Transactions on VLSI*
  - Q2 in Hardware Architecture
  - Impact Factor: ~2.312 (@2022)
  - Approx. 3.5 citations per paper
- Project: *PEPCC (Power Efficiency and Performance for Embedded and HPC Systems with Custom CGRAs)*
- Keywords: *Surveys and overviews; Hardware Accelerators*
  - Binary acceleration, instruction traces, (automated) hardware synthesis

Specific focus: reutilizing co-processor area via DPR to accelerate multiple loops w/o additional hardware cost

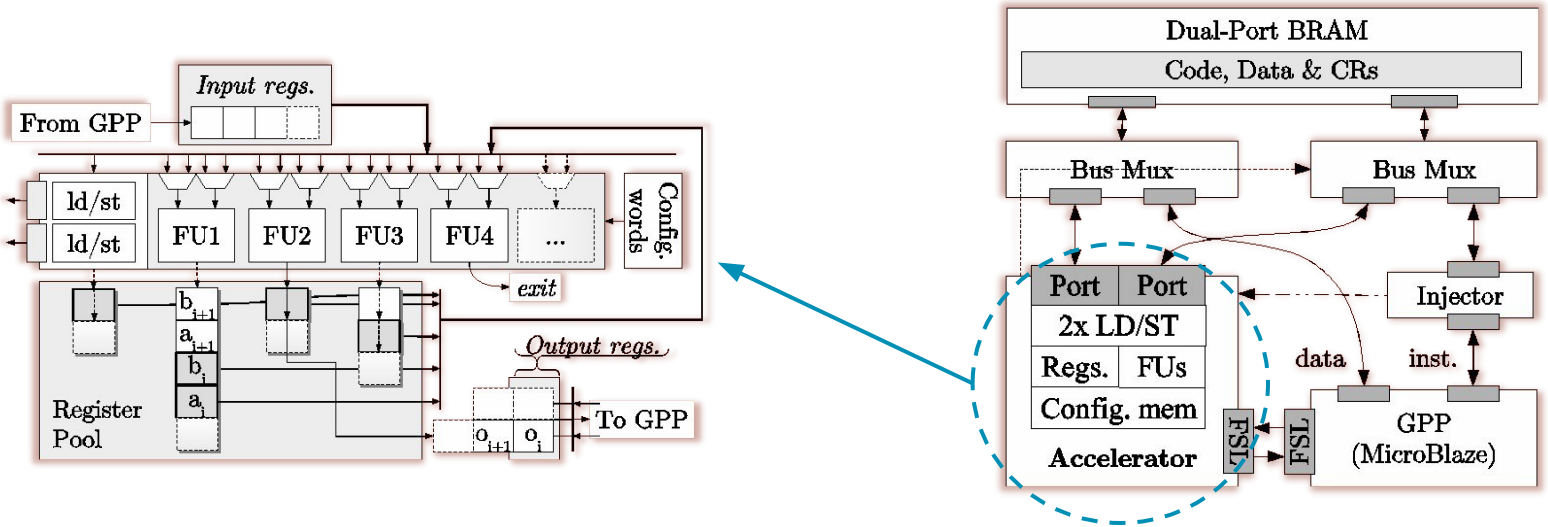
Previous Work - General Overview



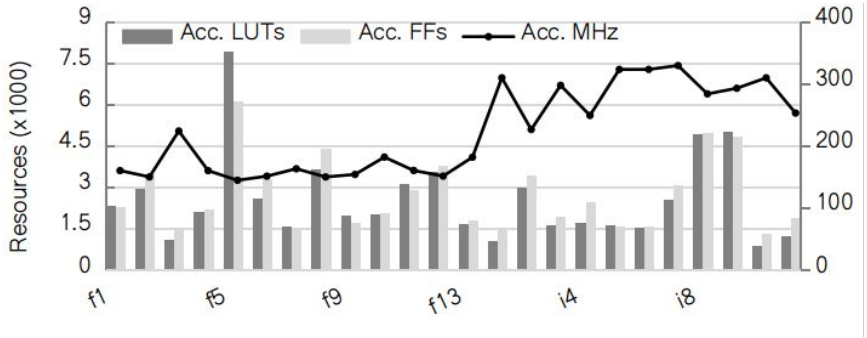
1. Identify frequent binary loop traces (existing work)
2. Translate loops into hardware accelerators
3. Detect imminent execution of loops at runtime
4. Migrate execution to accelerators

Previous Work - Architecture

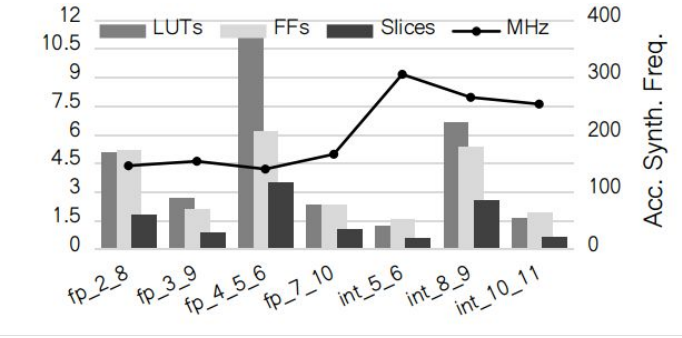
- 1 MicroBlaze + Loop accelerator (1+ configs); shared memory
- Loop accelerator: modulo-scheduled loops; specialized FUs & connections



Previous Work - Results



Resources for instances with a single loop scheduled

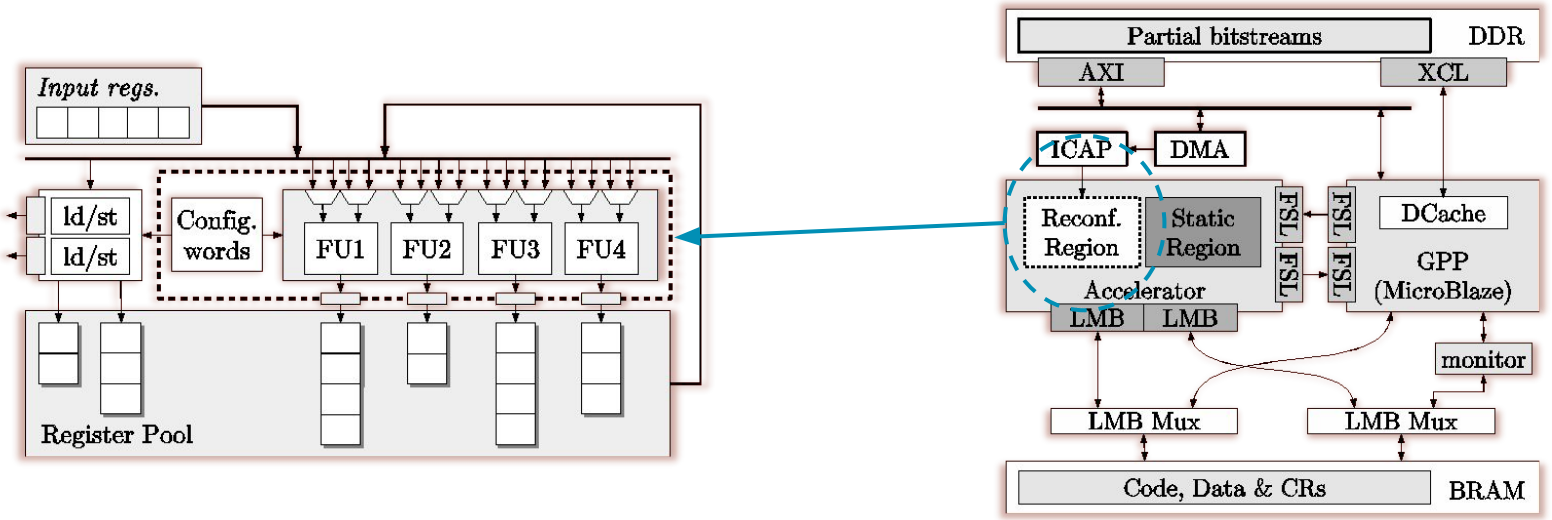


Resources for instances with a multiple loops scheduled

- Geometric mean speedups between 1.47x and 18.98x for 24 benchmarks
- Issue: for too many loop kernels → large area required & drops in frequency

Adding DPR

- DPR → Change a region of the FPGA configuration at runtime
  - **Reutilize resources** by changing the Functional Units, config memory, and muxes

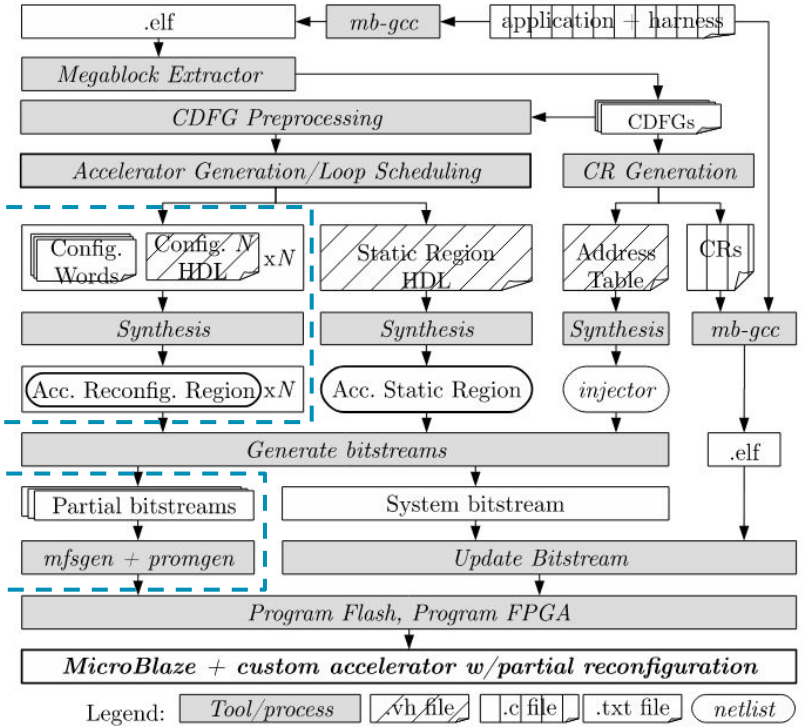


Toolflow Including Partial Bitstream Generation

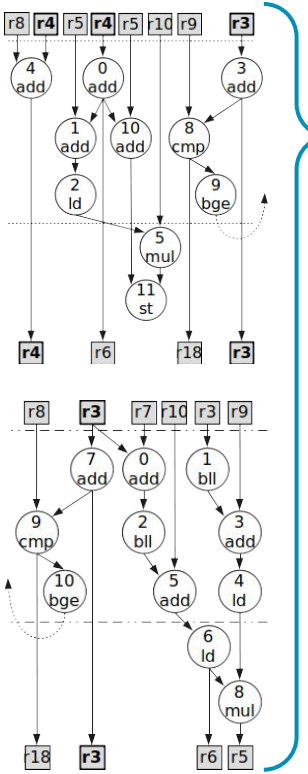
- Significant restructuring of tools and build flows…

- Up to N reconfigurable region variants
  - From: 1 Variant with N configs
  - To: N Variants with 1 Config

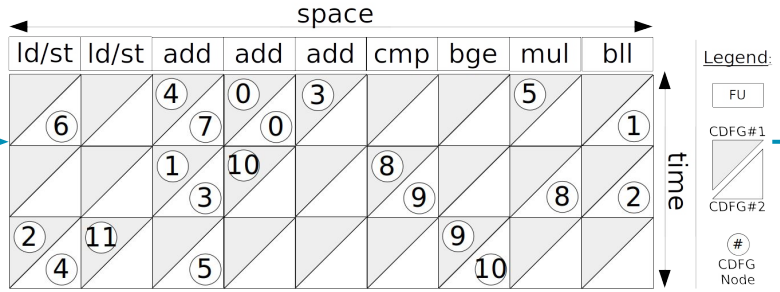
- File system with partial bitstream files



Toolflow - 1 or More CDFGs per Partial Bitstream

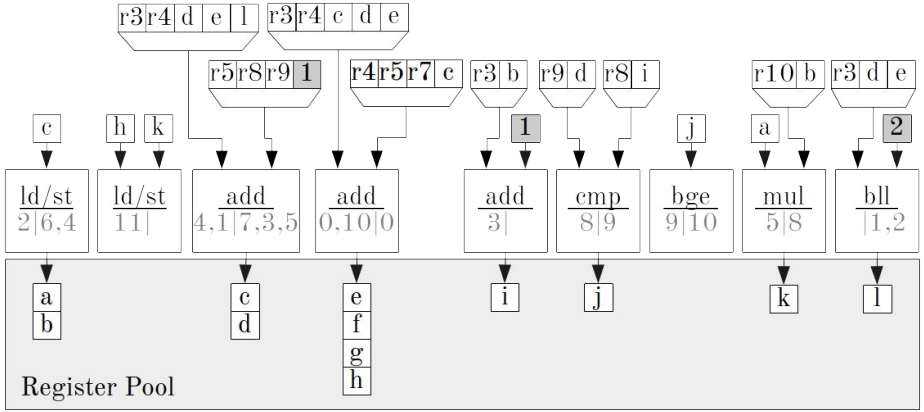


Schedule 2 CDFGs



Generate HDL

Instance with 2 configs held in 1 partial bitstream



Experimental Setup

- Xilinx VC707 Evaluation Board
  - Virtex-7 VX485T FPGA
- 21 integer and floating-point loop kernels
  - Livermore Loops
  - Texas Instruments IMGLIB

- Three sets of kernel groups
  - Small → 2 to 3 loops → 8 groups
  - Medium → 5 to 6 loops → 4 groups
  - Large → 10 to 11 loops → 2 groups
- For each group → generate CLA instances:
  - w/ **a single partial bitstream** → Faster switching; more area required
  - w/ **multiple bitstreams** → Area savings; frequency decrease less likely; DPR overhead

Experimental Setup

- Kernel groups
  - 3 group sizes
  - One instance per group
  - Average #insts: 35
  - Average II: 7 clock cyles

- All CDFGs scheduled at min. II regardless of use of DPR

- More configs. → more area *if* DPR is not used

kernel	#insts	II	small	medium	large
quantize	10	3	s1	m1	b1
perimeter	21	3			
boundary	24	8			
conv3x3	64	10	s2		
sad16	13	2			
mad16	13	2			
sobel	40	5	s3	m2	
dilate	142	29			
erode	142	29			
innerprod	10	3	s4		
matmul	15	3			
hydro	17	3	s5	m3	b2
hydro2d	37	6			
innprodfp	10	40	s6		
matmulfp	15	3			
intpredict	41	10	s7	m4	
diffpredict	44	10			
glinearrec	13	3			
cholesky	20	3	s8		
statefrag	42	5			
tridiag	12	3			

Experimental Results

Kernel Set	Avg. Resources within allocated area for **single partial bitstream** (all configs in 1 circuit)		
	Slices / (%)	LUTs / (%)	Impl. Time (m)
small set	1667 / 91%	5760 / 78%	93.7
medium set	2669 / 97%	9753 / 90%	44.6
big set	6395 / 83%	18855 / 62%	70.1

Kernel Set	Avg. Resources within allocated area for **multiple partial bitstreams** (1 circuit per config)		
	Slices / (%)	LUTs / (%)	Impl. Time (m)
small set	1195 / 84%	3825 / 69%	57.0
medium set	1660 / 91%	5139 / 70%	89.1
big set	2073 / 92%	6405 / 71%	162.6

- Area reserved for units, muxes, and configuration memory decreases considerably when resorting to DPR

Experimental Results

	Static Area			Reconfigurable Area		
	Slices	LUTs	FFs	Slices	LUTs	FFs
small set	0.40	0.16	0.87	1.32	3.25	0.05
medium set	0.77	0.19	1.55	2.01	5.50	0.07
big set	1.56	0.45	3.03	4.81	10.64	0.09

Resources for entire accelerator, for single partial bitstream, normalized to 1 MicroBlaze

	Static Area			Reconfigurable Area		
	Slices	LUTs	FFs	Slices	LUTs	FFs
small set	0.31	0.14	0.65	0.95	2.16	0.05
medium set	0.52	0.15	0.96	1.32	2.90	0.06
big set	0.63	0.19	1.17	1.62	3.61	0.08

Resources for entire accelerator, for multiple partial bitstreams, normalized to 1 MicroBlaze

Conclusion

- Mixed grain reconfiguration via DPR
- Up to 4.3x less area vs. larger circuit with larger config. memory depth
- Shortened synthesis times, despite multiple partial bitstreams
- More complex build flow
- Future work/issues
  - Reutilizing this idea for new system designs, e.g., RISC-V systems
  - Reconfigurable partitions unsupported in a design packed as an IP (with Xilinx tools)…